details of their post-translational modifications. 2-D gel databases are beginning to be linked to or integrated with comprehensive protein and nucleic acid databases (Neidhardt *et al.*, 1989; Simpson *et al.*, 1992; Appel *et al.*, 1994), and 'organism' databases, containing DNA sequence data, chromosomal map locations, reference 2-D gels and protein functional information for an organism, are becoming established as genome and proteome projects progress (VanBogelen *et al.*, 1992; Yeast Protein Database cited in Garrels *et al.*, 1994).

GEL IMAGE ANALYSIS AND REFERENCE GELS

After 2-D electrophoresis and protein visualisation by staining, fluorography or phosphorimaging, images of gels are digitised for computer analysis by an image scanner, laser densitomer, or charge-coupled device (CCD) camera (Garrels, 1989; Celis *et al.*, 1990a; Urwin and Jackson, 1993). All systems digitise gels with a resolution of 100 – 200 mm, and can detect a wide range of densities or shading (256 or more 'grey scales'). Following this, gel images are subjected to a series of manipulations to remove vertical and horizontal streaking and background haze, to detect spot positions and boundaries, and to calculate spot intensity (*Figure 3*). A standard spot (SSP) number, containing vertical and horizontal positional information, is assigned to each detected spot and becomes the protein's reference number. *Table 2* lists some notable software packages which process 2-D gel images.

Table 2: Some Software Packages for the Analysis of Gel Images.

Gel Image Analysis System	References*
ELSIE 4 & 5	Olsen and Miller, 1988; Wirth *et al.*, 1991; Wirth *et al.*, 1993.
GELLAB I & II	Wu, Lemkin and Upton, 1993; Lemkin, Wu and Upton, 1993; Myrick *et al.*, 1993.
MELANIE I & II	Appel, *et al.* 1991; Hochstrasser *et al.* 1991b.
QUEST I & II and PDQUEST	Garrels, 1989; Monardo *et al.*, 1994; Holt *et al.*, 1992; Celis *et al.*, 1990a,b.
TYCHO & KEPLAR	Anderson *et al.*, 1984; Richardson, Horn and Anderson, 1994.

* These references are not exhaustive; they include some references of use as well as authors of the system.

As there are difficulties in the electrophoresis of samples with 100% reproducibility, reference gel images are often constructed from many gels of the same sample (Garrels and Franza, 1989; Neidhardt *et al.*, 1989). Since this involves the matching of 2000 to 4000 proteins from one gel to another, it presents a considerable challenge to image analysis systems. Matching of gels is usually initiated by an operator, who manually designates approximately 50 or so prominent spots as 'landmarks' on gels to be cross-matched. Proteins which match are then established around landmarks, using computer-based vector algorithms to extend the matching over the entire gel. Close to 100% of spots from complex samples can be matched by these methods, although different degrees of operator intervention may be required (Olsen and Miller, 1988; Lemkin and Lester, 1989; Garrels, 1989; Myrick *et al.*, 1993).

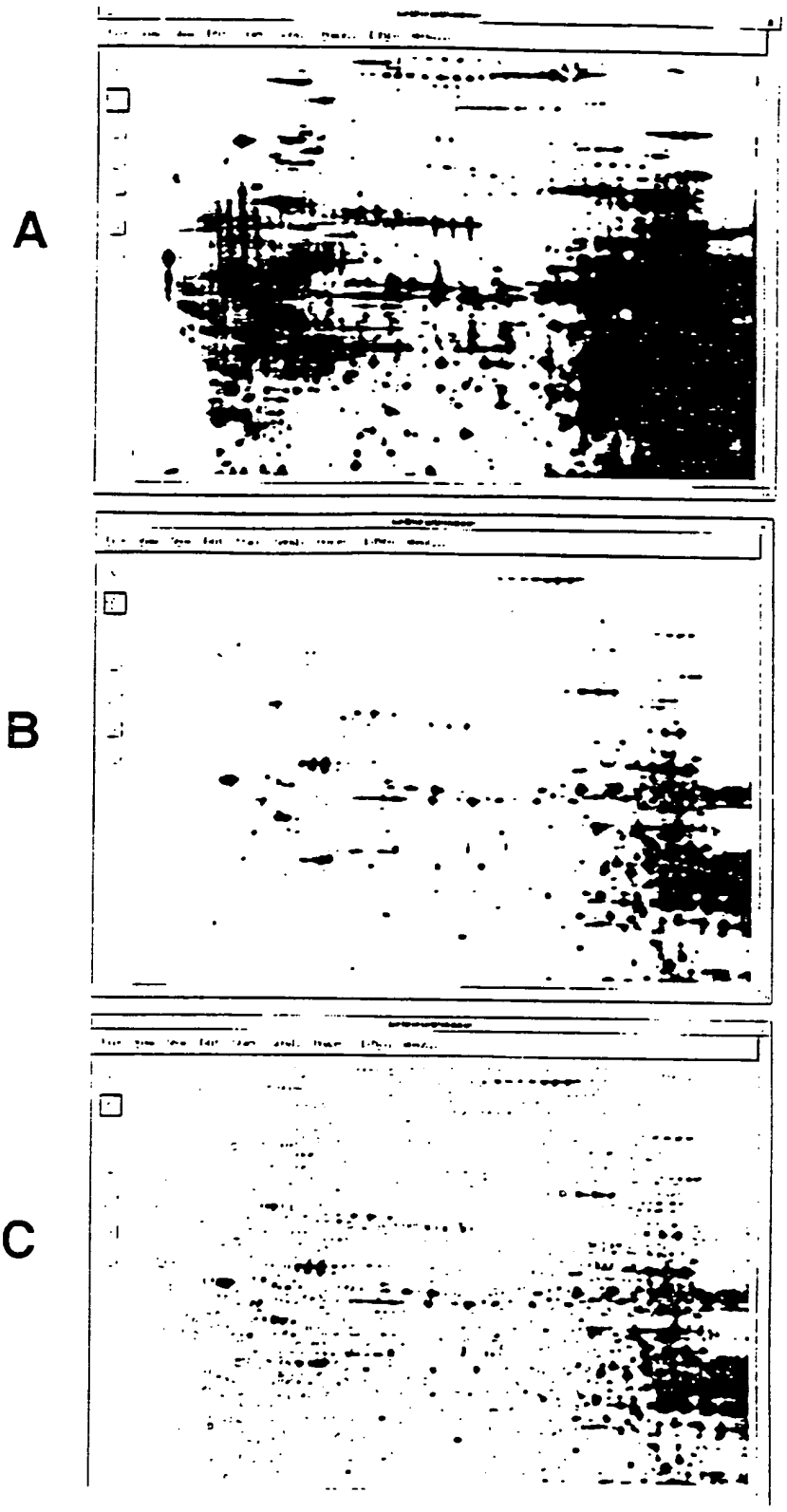

Figure 3. Computer processing of gel images. Shown is a wide pI range 2-D separation of human liver proteins, processed by Melanie software (Appel *et al.*, 1991). (A) Original gel image as captured by laser densitometer. (B) Gel image after processing to remove streaking and background. (C) Outline definition of all spots on the gel.

CALCULATION OF PROTEIN ISOELECTRIC POINT AND MOLECULAR WEIGHT

Estimation of the isoelectric point (pI) and molecular weight (MW) of proteins from 2-D gels provides fundamental parameters for each protein, which are also of use during identification procedures (see following section). The pI and MW of proteins are recorded in 2-D gel databases. Accurate estimations of protein pI and MW can be obtained by using 20 or more known proteins on a reference map to construct standard curves of pI and molecular weight, which are then used to calculate estimated pI and MW of unknown proteins (Neidhardt *et al.*, 1989; Garrels and Franza, 1989; Van-Bogelen, Hutton and Neidhardt, 1990; Anderson and Anderson, 1991; Anderson *et al.*, 1991; Latham *et al.*, 1992). Alternatively, the MW of individual proteins blotted to PVDF can be determined very accurately by direct mass spectrometry (Eckerskorn *et al.*, 1992). Where immobilised pH gradients are used, the focusing position of proteins allows their pI to be measured within 0.15 units of that calculated from the amino acid sequence (Bjellqvist *et al.*, 1993c). It must be noted, however, that proteins carrying post-translational modifications may migrate to unexpected pI or MW positions during electrophoresis (Packer *et al.*, 1995).

SPOT QUANTITATION AND EXPRESSION ANALYSIS

A major challenge faced in proteome projects is the quantitative analysis of proteins separated by 2-D electrophoresis. The most accurate means of protein quantitation is to determine chemically the amount of each protein present by amino acid compositional analysis. However, the current method of choice for quantitative analysis of many proteins is to radiolabel samples with [^{35}S] methionine or ^{14}C amino acids, perform the 2-D electrophoresis, and measure protein levels in disintegrations per minute (dpm) or units of optical density. Quantitation is achieved either by liquid scintillation counting, or by gel image analysis where spot densities are quantitated by reference to gel calibration strips containing known amounts of radiolabelled protein or against the integrated optical density of all spots visualised (Vandekerkhove *et al.*, 1990; Celis *et al.*, 1990b; Celis and Olsen, 1994; Garrels, 1989; Latham, Garrels and Solter, 1993; Fey *et al.*, 1994). All approaches effectively allow spots to be normalised against the total disintegrations per minute loaded onto the gel. Limitations that remain with radiolabelling methods are that absolute quantitation is not achieved because all proteins have varying amounts of any amino acid, and that only easily labelled samples can be investigated. Quantitative silver staining presents an alternative (Giometti *et al.*, 1991; Harrington *et al.*, 1992; Rodriguez *et al.*, 1993; Myrick *et al.*, 1993), which when undertaken with [^{35}S]thiourea (Wallace and Saluz, 1992 a,b) is of extremely high sensitivity.

When protein spots from samples prepared under different conditions are quantitated and matched from gel to gel, it becomes possible to examine changes and patterns in protein expression. Large scale investigation of up- and down-regulation of proteins, their appearance and disappearance, can be undertaken. For example, simian virus 40 transformed human keratinocytes were shown to have 177 up-regulated and 58 down-regulated proteins compared to normal keratinocytes (Celis and Olsen, 1994); detailed synthesis profiles of 1200 proteins have been established in 1 to 4 cell mouse embryos (Latham *et al.*, 1991, 1992); and 4 proteins out of 1971 were found to be markers for

cadmium toxicity in urinary proteins (Myrick *et al.*, 1993). Complex global changes in protein expression as a result of gene disruptions have also been investigated (S. Fey and P. Mose-Larsen, Personal communication). Impressively, large gel sets showing protein expression under different conditions can be globally investigated using statistical methods that find groups of related objects within a set. For example, the REF52 rat cell line database, consisting of 79 gels from 12 experimental groups where each gel contains quantitative data for 1600 cross-matched proteins, has been analysed by cluster analysis (Garrels *et al.*, 1990). This revealed clusters of proteins that, for example, were induced or repressed similarly under simian virus 40 or adenovirus transformation, suggesting a common mechanism. Protein groups that were induced or repressed during culture growth to confluence were also found. It is obvious that the potential for investigation of cellular control mechanisms by these approaches is immense. It is equally clear that investigations of gene expression of this scale are currently technically impossible using nucleic-acid based techniques.

Table 3: Some proteome databases and their special features

Proteome database	Special features	References
E. coli gene-protein database	Gel spots linked with GenBank and Kohara clones; quantitative spot measurements under different growth conditions	VanBogelen and Neidhardt, 1991; VanBogelen *et al.*, 1992
Human heart databases	Identification of disease markers; two separate databases have been established	Baker *et al.*, 1992 Corbett *et al.*, 1994b Junghlut *et al.*, 1994
Human keratinocyte database	Extensive identifications; quantitative spot measurements of transformed cells; identification of disease markers	Celis *et al.*, 1990a Celis *et al.*, 1993 Celis and Olsen 1994
Mouse embryo database	Quantitative spot measurements through 1 to 4 cell stage	Latham *et al.*, 1991 Latham *et al.*, 1992
Mouse liver database (Argonne Protein Mapping Group)	Documents changes due to exposure to ionizing radiation and toxic chemicals	Giometti, Taylor and Tollaksen, 1992
Rat liver epithelial database	Detailed subcellular fractionation studies	Wirth *et al.*, 1991 Wirth *et al.*, 1993
Rat liver database	Extensive studies on regulation of proteins by drugs and toxic agents	Anderson and Anderson, 1991; Anderson *et al.*, 1992; Richardson, Horn and Anderson, 1994
REF 52 rat cell line database	Accessible via World Wide Web; quantitative spot measurements under different conditions	Garrels and Franza 1989 Boutell *et al.*, 1994
SWISS-2DPAGE containing human reference maps	Accessible via World Wide Web; completely integrated with SWISS-PROT and SWISS-3DIMAGE	Appel *et al.*, 1993 Hochstrasser *et al.*, 1992 Hughes *et al.*, 1993 Golaz *et al.*, 1993
Yeast Protein Database (YPD) and Yeast Electrophoretic Protein Database (YEPD)	Completely crossreferenced organism database; YPD has extensive information on over 3500 proteins; YEPD has many identifications	Garrels *et al.*, 1994

FEATURES OF PROTEOME DATABASES

Proteome projects rely heavily on computer databases to store information about all proteins expressed by an organism. 'Proteome databases' should contain detailed information of proteins already characterised elsewhere, as well as protein data from 2-D gels such as apparent pI and MW, expression level under different conditions, subcellular localisation, and information on post-translational modifications. Images of reference 2-D gels, showing protein SSP numbers and protein identifications, should also be included. Ideally, proteome databases should be accessible with Macintosh or IBM personal computers and easy to use. Some proteome databases and the areas they cover are listed in *Table 3*. Databases range from collections of annotated gels to large databases of images integrated with protein and nucleic acid sequence banks.

One example of an integrated proteome database is the suite of SWISS-PROT, SWISS-2DPAGE and SWISS-3DIMAGE databases (Appel *et al.*, 1993; Appel *et al.*, 1994; Appel, Bairoch and Hochstrasser, 1994; Bairoch and Boeckmann, 1994). The features of these three databases are listed in *Table 4*. SWISS-PROT, SWISS-2DPAGE and SWISS-3DIMAGE are accessible through the World Wide Web

Table 4: The SWISS-PROT, SWISS-2DPAGE and SWISS-3DIMAGE suite of crosslinked databases. All three databases are accessible through the World Wide Web, at URL address: http://expasy.hcuge.ch/

	SWISS-PROT	SWISS-2DPAGE	SWISS-3DIMAGE
Information	Text entries of sequence data; Citation information; taxonomic data; 38, 303 entries in Release 29	2-D gel images of: human liver, plasma, HepG2, HepG2 secreted proteins, red blood cell, lymphoma, cerebrospinal fluid, macrophage like cell line, erythroleukemia cell, platelet	Collection of 330 3-D images of proteins
Annotations	Protein function; Post translational modifications; Domains; Secondary structure; Quaternary structure; Diseases associated with protein; Sequence conflicts	Gel images where protein is found; How protein identified; Protein pI and MW; protein number; normal and pathological variants	All annotation is available in SWISS-PROT
Cross-Referenced Databases	SWISS-2DPAGE SWISS-3DIMAGE EMBL; PIR; PDB; OMIM; PROSITE; Medline; Flybase; GCRDb; MaizeDB; WormPep; DictyDB	SWISS-PROT and all other databases accessible through SWISS-PROT	SWISS-PROT and all other databases accessible through SWISS-PROT
Other Features	Navigation to other SWISS databases achieved by selecting entries with computer mouse	Gel images show position of identified proteins, or region of gel where protein should appear	Mono and stereo images available; Images can be transferred to local computer image viewing programs

(Berners-Lee *et al.*, 1992), allowing any computer connected to the internet to access the stored information and images. Navigation within and between the three databases is seamless, as all potential crosslinks are highlighted as hypertext on the display and can be selected with a computer mouse. From these databases, detailed information about a protein, including amino acid sequence and known post-translational modifications, can be obtained, the precise protein spot it corresponds to on a reference gel image can be viewed if known, and the 3-D structure of the molecule can be seen if available. References to nucleic acid and other databases are also given to provide access to information stored elsewhere.

Organism databases, containing detailed protein and nucleic acid information about a species, are becoming common as genome and proteome projects progress. These differ from nucleic acid or protein sequence databases like GenBank or SWISS-PROT because they are image based, and contain information about chromosomal map positions, transcription of genes, and protein expression patterns. The *Escherichia coli* gene-protein database (VanBogelen, Hutton and Neidhardt, 1990; VanBogelen and Neidhardt, 1991; VanBogelen *et al.*, 1992), known as the ECO2DBASE, is one example. It contains gene and protein names, 2-D gel spot information (including pI and MW estimates, and spot identification), genetic information (GenBank or EMBL codes, chromosomal location, location on Kohara clones (Kohara, Akiyama, and Isono, 1987), transcription direction of genes), and protein regulatory information (level of protein expression under different growth regimes, member of regulon or stimulon). All entries in the ECO2DBASE are also cross-referenced to the SWISS-PROT database (Bairoch and Boeckmann, 1994). It is anticipated that organism databases will soon become a standard means of storing all available information about a particular species. However there is currently no consistent manner in which organism databases are assembled, which may hamper comparisons in the future.

Identification and characterisation of proteins from 2-D gels

The number of proteins identified on a 2-D reference map determines its usefulness as a research and reference tool. As most reference maps have only a small proportion of proteins identified, a major aim of current proteome projects is to screen many proteins from 2-D maps, in order to define them as 'known' in current nucleic acid and protein databases, or as 'unknown'. Protein identification assists in confirmation of DNA open reading frames, and provides focus for DNA sequencing projects and protein characterisation efforts by pointing to proteins that are novel. Since there may be 3000–4000 proteins from a single 2-D map that require identification, the challenge in protein screening is to identify proteins quickly, with a minimum of cost and effort.

Traditionally, proteins from 2-D gels have been identified by techniques such as immunoblotting, N-terminal microsequencing, internal peptide sequencing, comigration of unknown proteins with known proteins, or by overexpression of homologous genes of interest in the organism under study (Matsudaira, 1987; Rosenfeld *et al.*, 1992; VanBogelen *et al.*, 1992; Celis *et al.*, 1993; Honore *et al.*, 1993; Garrels *et al.*, 1994). Whilst these techniques are powerful identification tools, they are too expensive or time and labour intensive to use in mass screening programs. A hierarchical approach to mass protein identification has been recently suggested as an

Table 5: Hierarchical analysis for mass screening of 2-D separated proteins blotted to membranes. Rapid and inexpensive techniques are used as a first step in protein identification, and slower, more expensive techniques are then used if necessary. Table modified from Wasinger *et al.*, 1995).

Order	Identification technique	References
1	Amino acid analysis	Junghlut *et al.*, 1992; Shaw, 1993; Hobohm, Houthaeve and Sander, 1994; Junghlut *et al.*, 1994; Wilkins *et al.*, 1995
2	Amino acid analysis with N-terminal sequence tag	Wilkins *et al.*, submitted
3	Peptide-mass fingerprinting	Henzel *et al.*, 1993; Pappin, Hojrup and Bleasby, 1993; James *et al.*, 1993; Mann, Hojrup and Roepstorff, 1993; Yates *et al.*, 1993; Moritz *et al.*, 1994; Sutton *et al.*, 1995
4	Combination of amino acid analysis and peptide mass fingerprinting	Cordwell *et al.*, 1995; Wasinger *et al.*, 1995;
5	Mass spectrometry sequence tag	Mann and Wilm, 1994
6	Extensive N-terminal Edman microsequencing	Matsudaira, 1987
7	Internal peptide Edman microsequencing	Rosenfeld *et al.*, 1992; Hellman *et al.*, 1995;
8	Microsequencing by mass spectrometry (electrospray ionisation, post-source decay MALDI-TOF)	Johnson and Walsh, 1992
9	Ladder sequencing	Bartlet-Jones *et al.*, 1994

alternative to traditional approaches (*Table 5*; Wasinger *et al.*, 1995). This involves the use of rapid and cheap identification tools such as amino acid analysis and peptide mass fingerprinting as first steps in protein identification, followed by the use of slower, more expensive and time consuming identification procedures if necessary. In the construction of this hierarchy the analysis time, cost per sample and the complexity of the data created has been considered, as whilst some techniques require little machine time per sample, the analysis of data can be quite involved and time consuming. Amino acid analysis and peptide mass-fingerprinting based identification techniques in the hierarchy are discussed in detail below. For review of other protein identification techniques in *Table 5*, see Patterson (1994) and Mann (1995).

PROTEIN IDENTIFICATION BY AMINO ACID COMPOSITION

There has been a revival of interest in the use of amino acid composition for identification of proteins from 2-D gels after early work by Eckerskorn *et al.* (1988). This technique uses a protein's idiosyncratic amino acid composition profile in order to identify it by comparison with theoretical compositions of proteins in databases. The amino acid composition of proteins can be determined by differential metabolic radiolabelling and quantitative autoradiography after 2-D electrophoresis (Garrels *et al.*, 1994; Frey *et al.*, 1994), or by acid hydrolysis of membrane-blotted proteins and chromatographic analysis of the resulting amino acid mixture (Eckerskorn *et al.*, 1988; Tous *et al.*, 1989; Gharahdaghi *et al.*, 1992; Jungblut *et al.*, 1992; Wilkins *et al.*, 1995). As differential metabolic labelling experiments require X-ray film or phosphor-image plate exposures of up to 140 days, and can only be undertaken with easily radiolabelled samples, the technique is not as rapid or widely applicable as chromato-

```
Spot ECOLI-B1M
==============

Composition:

Asx: 13.2   Glx: 10.4   Ser:  5.7   His:  0.7
Gly:  5.4   Thr:  3.8   Ala:  6.7   Pro:  7.9
Tyr:  1.3   Arg:  5.0   Val:  8.0   Met:  0.3
Ile:  5.9   Leu:  8.0   Phe: 13.3   Lys:  4.4

pI estimate:   6.89  Range searched: ( 6.64,  7.14)
Mw estimate:  16800  Range searched: (13440, 20160)

Closest SWISS-PROT entries for the species ECOLI matched by AA composition:

Rank Score   Protein     pI      Mw     Description
=============================================
   1   24  PYRI_ECOLI   6.84    16989  ASPARTATE CARBAMOYLTRANSFERASE
   2   39  COAA_ECOLI   6.32    36359  PANTOTHENATE KINASE (EC 2.7.1.33)
   3   40  META_ECOLI   5.06    35713  HOMOSERINE O-SUCCINYLTRANSFERASE
   4   42  CADC_ECOLI   5.52    57812  TRANSCRIPTIONAL ACTIVATOR CADC.
   5   43  HLYC_ECOLI   8.58    19769  HEMOLYSIN C, PLASMID.

Closest SWISS-PROT entries for ECOLI with pI and Mw values in specified
range:

Rank Score   Protein     pI      Mw     Description
=============================================
   1   24  PYRI_ECOLI   6.84    16989  ASPARTATE CARBAMOYLTRANSFERASE
   2  102  TRJ8_ECOLI   6.73    17921  TRAJ PROTEIN.
   3  112  YAJG_ECOLI   6.79    19028  HYPOTHETICAL LIPOPROTEIN YAJG.
   4  140  YFJB_ECOLI   6.83    14945  HYPOTHETICAL 14.9 KD PROTEIN IN GRPE
   5  142  YAHA_ECOLI   7.06    14726  HYPOTHETICAL PROTEIN IN BETT 3'REGION
```

Figure 4. Computer printout from ExPASy server where the empirical amino acid composition, estimated pI and MW of a protein from a 2-D reference map of *E. coli* were matched against all entries in SWISS-PROT for *E. coli*. The correct identification, aspartate carbamoyltransferase, is shown in bold. Low scores indicate a good match. Note how matching within a defined pI and MW range (lower set of proteins) has greatly increased the score difference between the first and second ranking proteins. This score difference gives high confidence in the identification, and is only observed where the top ranking protein is the correct identification (Wilkins *et al.*, 1995).

graphy-based analysis. Proteins blotted to PVDF membranes can be hydrolysed in 1 h at 155°C, amino acids extracted in a single brief step, and each sample automatically derivatised and separated by chromatography in under 40 minutes (Wilkins *et al.*, 1995; Ou *et al.*, 1995). In this manner, one operator can routinely analyse 100 proteins per week on one HPLC unit. This technology lends itself to automation, and it is anticipated that instruments with even greater sample throughput will be developed. When proteins have been prepared by micropreparative 2-D electrophoresis (Hanash *et al.*, 1991; Bjellqvist *et al.*, 1993b), blotted to a PVDF membrane and stained with amido black, any visible protein spot is of sufficient quantity for amino acid analysis (Cordwell *et al.*, 1995; Wasinger *et al.*, 1995; Wilkins *et al.*, 1995).

After the amino acid composition of a protein has been determined, computer programs are used to match it against the calculated compositions of proteins in databases (Eckerskorn *et al.*, 1988; Sibbald, Sommerfeldt and Argos, 1991; Jungblut *et al.*, 1992; Shaw, 1993; Hobohm, Houthaeve and Sander, 1994; Wilkins *et al.*, 1995). Matching is usually done with only 15 or 16 amino acids, as cysteine and

```
Spot ECOLI-ACJ
==============

Composition:

Asx:  9.6   Glx: 10.8   Ser:  4.1   His:  2.7
Gly: 12.2   Thr:  3.8   Ala: 11.9   Pro:  3.2
Tyr:  6.0   Arg:  3.7   Val:  9.5   Met:  0.6
Ile:  5.0   Leu:  8.2   Phe:  3.2   Lys:  4.9

pI estimate:   5.99  Range searched: ( 5.74,  6.24)
Mw estimate:  45000  Range searched: (36000, 54000)

Closest SWISS-PROT entries for ECOLI with pI and Mw values in specified
range:

Rank Score   Protein    pI       Mw    N-terminal Seq.
=============================================================
   1   21  GLYA_ECOLI   6.03    45316    M L K R E
   2   32  YJGB_ECOLI   5.86    36502    M S M I K
   3   38  GABT_ECOLI   5.78    45774    M S N S K
   4   44  YIHS_ECOLI   5.86    48018    M R I K Y
   5   45  DHE4_ECOLI   5.98    48581    M D Q T Y
   6   46  ARGD_ECOLI   5.79    43765    M A I E Q
   7   46  MJRB_ECOLI   5.78    37851    M N H S L
   8   47  GLMU_ECOLI   5.98    49162    M L N N A
   9   47  ACKA_ECOLI   5.85    43290    M S S K L
  10   50  YJJN_ECOLI   6.01    37064    M E S R I
```

Figure 5. A PVDF protein spot from an *E. coli* 2-D reference map was sequenced for 4 cycles, and the same sample then subject to amino acid analysis. The N-terminal sequence was M L K R. When the amino acid composition of the spot, as well as estimated pI and MW, were matched against all entries in SWISS-PROT for *E. coli*, the above list of best matches was produced. N-terminal sequences are from SWISS-PROT for those entries. The top ranking identification of serine hydroxymethyltransferase (bold) did not show a large score difference between the first and second ranking proteins, giving little confidence in this being the correct protein identification. However, the sequence tag (M L K R) confirmed the identity of the protein as serine hydroxymethyltransferase.

tryptophan are destroyed during hydrolysis, asparagine and glutamine are deamidated to their corresponding acids, and proline is not quantitated in some analysis systems. The computer programs produce a list of best matching proteins, which are ranked by a score that indicates the match quality. Some programs allow matching to be restricted to specific 'windows' of MW and pI (Hobohm, Houthaeve and Sander, 1994; Wilkins *et al.*, 1995), and to protein database entries for one species (Jungblut *et al.*, 1992; Wilkins *et al.*, 1995). The use of such restrictions increases the power of matching. An example of protein identification by amino acid composition is shown in *Figure 4*. To date, amino acid composition has been used to identify proteins from reference maps of *Spiroplasma melliferum*, *Mycoplasma genitalium*, *E. coli*, *Saccharomyces cerevisiae*, *Dictyostelium discoideum*, human sera, human heart, human lymphocyte, and mouse brain (Cordwell *et al.*, 1995; Wasinger *et al.*, 1995; Wilkins *et al.*, 1995; Jungblut *et al.*, 1992, 1994; Garrels *et al.*, 1994; Frey *et al.*, 1994).

PROTEIN IDENTIFICATION BY AMINO ACID COMPOSITION AND N-TERMINAL SEQUENCE TAG

When samples from 2-D gels are not unambiguously identified by amino acid

composition, pI and MW, often the correct identification of that protein is amongst the top rankings of the list (Hobohm, Houthaeve and Sander, 1994; Cordwell *et al.*, 1995; Wilkins *et al.*, 1995). Taking advantage of this observation, we have used the mass spectrometry 'sequence tag' concept (Mann and Wilm, 1994) in developing a combined Edman degradation and amino acid analysis approach to protein identification (Wilkins *et al.*, submitted). This involves the N-terminal sequencing of PVDF-blotted proteins by Edman degradation for 3 or 4 cycles to create a 'sequence tag', following which the same sample is used for amino acid analysis. As only a few amino acids are removed from the protein, its composition is not significantly altered. Furthermore, since only a small amount of protein sequence is required, fast but low repetitive yield Edman degradation cycles can be used. Modifications to current procedures should allow 3 cycles to be completed in 1 h, thereby allowing the screening of 100 or more proteins per week on one automated, multi-cartridge sequenator. Amino acid composition, pI and MW of proteins are matched against databases as described above, and N-terminal sequences of best matching proteins are checked with the 'sequence tag' to confirm the protein identity (*Figure 5*). This technique will be less useful when proteins are N-terminally blocked, but as only a few N-terminal amino acids are susceptible to the acetyl, formyl, or pyroglutamyl modifications that cause blockage, this may itself provide useful information for sequence tag identification. A strength of N-terminal sequence tag and amino acid composition protein identification is that data generated are quickly and easily interpreted.

PROTEIN IDENTIFICATION BY PEPTIDE MASS FINGERPRINTING

Techniques for the identification of proteins by peptide mass fingerprinting have recently been described (Henzel *et al.*, 1993; Pappin, Hojrup and Bleasby, 1993; James *et al.*, 1993; Mann, Hojrup and Roepstorff, 1993; Yates *et al.*, 1993; Mortz *et al.*, 1994; Sutton *et al.*, 1995). This involves the generation of peptides from proteins using residue-specific enzymes, the determination of peptide masses, and the matching of these masses against theoretical peptide libraries generated from protein sequence databases. As proteins have different amino acid sequences, their peptides should produce characteristic 'fingerprints'.

The first step of peptide mass fingerprinting is protein digestion. Proteins within the gel matrix or bound to PVDF can be enzymatically digested *in situ*, although *in situ* gel digests are reported to produce more enzyme autodigestion products, which complicate subsequent peptide mass analysis (James *et al.*, 1993; Rasmussen *et al.*, 1994; Mortz *et al.*, 1994). The enzyme of choice for digestion is currently trypsin (of modified sequencing grade), but other enzymes (Lys-C or *S. aureus* V8 protease) have also been used (Pappin, Hojrup and Bleasby, 1993). To maximise the number of peptides obtained, it is desirable for protein samples to be reduced and alkylated prior to digestion (Mortz *et al.*, 1994; Henzel *et al.*, 1993). This ensures that all disulfide bonds of the protein are broken, and produces protein conformations that are more amenable to digestion. Surprisingly, chemical digestion methods such as cyanogen bromide (methionine specific), formic acid (aspartic acid specific), and 2-(2'-nitrophenylsulfenyl)-3-methyl-3'-bromoindolenine (tryptophan specific) have not been explored as means of peptide production for mass fingerprinting, even though they are rapid and may circumvent some problems associated with enzyme digestions

(Nikodem and Fresco, 1979; Crimmins *et al.*, 1990; Vanfleteren *et al.*, 1992).

After proteins are digested, peptide masses are determined by mass spectrometry. Direct analysis of peptide mixtures can be achieved by electrospray ionisation mass spectrometry, plasma desorption mass spectrometry, or matrix assisted laser desorption ionization (MALDI) mass spectrometry techniques. MALDI is preferable because of its higher sensitivity and greater tolerance to contaminating substances from 2-D gels (James *et al.*, 1993; Mortz *et al.*, 1994; Pappin, Hojrup and Bleasby, 1993). Furthermore, recent modifications to sample preparation methods have largely solved early difficulties experienced with the calibration of MALDI spectra (Mortz *et al.*, 1994; Vorm and Mann, 1994; Vorm, Roepstorff and Mann, 1994). The high sensitivity of mass spectrometry allows a small fraction of a digest of a 1μg protein spot to be used for analysis, and analysis itself is complete in a few minutes.

A major challenge associated with peptide mass fingerprinting is data interpretation prior to computer matching against libraries of theoretical peptide digests. Spectra must be examined carefully to determine which peaks represent peptide masses of interest, as there are often enzyme autodigestion products and contaminating substances present (Henzel *et al.*, 1993; Mortz *et al.*, 1994; Rasmussen *et al.*, 1994). Furthermore, if protein alkylation and reduction has not been undertaken prior to protein digestion, peptide sequence coverage may be poor (40% to 70%), with some masses present representing disulfide bonded peptides originally present in the protein (Mortz *et al.*, 1994). For eukaryotes, a serious issue is the alteration of peptide masses by the presence of post-translational modifications (*Table 6*). The mass of the unmodified peptide alone can be very difficult to determine. Two artifactual modifications introduced by electrophoresis, an acrylamide adduct to cysteine and the oxidation of methionine, are also known to alter peptide masses (le Maire *et al.*, 1993; Hess *et al.*, 1993).

Table 6: Masses of some common post-translational modifications. Peptides carrying post-translational modifications complicate data analysis for peptide mass fingerprinting protein identification. This is especially so for protein glycosylation, which involves many different combinations of the hexosamines, hexoses, deoxyhexoses, and sialic acid

Post-translational modification	Mass change
Acetylation	+ 42.04
*Acrylamide adduct to cysteine	+71.00
Carboxylation of Asp or Glu	+ 44.01
Deamidation of Asn or Gln	+ 0.98
Disulfide bond formation	− 2.02
Deoxyhexoses (Fuc)	146.14
Formylation	+ 28.01
Hexosamines (GlcN, GalN)	+ 161.16
Hexoses (Glc, Gal, Man)	+ 162.14
Hydroxylation	+ 16.00
N-acetylhexosamines (GlcNAc, GalNAc)	+ 203.19
*Oxidation of Met	+ 16.00
Phosphorylation	+ 79.98
Pyroglutamic acid formed from Gln	−17.03
Sialic acid (NeuNAc)	+ 291.26
Sulfation	+ 80.06

Table modified from Finnigan LASERMAT application data sheet 5.
Asterisk * shows modifications that can arise artifactually from the 2-D electrophoresis process.

A number of computer programs are available for matching peptide masses against databases (reviewed in Cottrell, 1994). Matching is usually undertaken in an interactive manner, whereby peaks of mass 500–3000 Da are selected and matched under various search parameters including MW of protein, mass accuracy of peptides, and number of missed enzyme cleavages allowed (Henzel *et al.*, 1993; Mortz *et al.*, 1994; Rasmussen *et al.*, 1994). The correct protein identity is the protein which has the most peptide masses in common with the unknown sample. Identities have been established with as few as three peptides, but unambiguous identification is thought to require a mass spectrometric map covering most peptides of the protein (Mortz *et al.*, 1994; Yates *et al.*, 1993). To date, peptide mass fingerprinting of proteins has been undertaken from the human myocardial protein and keratinocyte maps, from an *E. coli* 2-D gel, and from reference maps of *Spiroplasma melliferum* and *Mycoplasma genitalium* (Sutton *et al.*, 1995; Rasmussen *et al.*, 1994; Henzel *et al.*, 1993; Cordwell *et al.*, 1995, Wasinger *et al.*, 1995), although the technique is most powerful when used in combination with another protein identification technique (Rasmussen *et al.*, 1994; Cordwell *et al.*, 1995).

MASS SPECTROMETRY SEQUENCE TAGGING

An extension of peptide mass fingerprinting has recently been described, called peptide sequence tagging (Mann and Wilm, 1994; Mann, 1995). This uses tandem mass spectrometry (MS/MS) to initially determine the mass of peptides, then subject them to fragmentation by collision with a gas, and finally determine the mass of fragments. The resulting spectra gives information about a peptide's amino acid sequence. The fragmentation masses of peptides can rarely be used to assign a complete sequence, but it usually allows a short 'sequence tag' of 2 or 3 amino acids to be determined. This sequence tag and the original peptide mass is matched by computer against a database, providing a likely identity of the peptide and the protein it came from. The major drawback for this technique as a mass screening tool is the complexity of the mass data generated and the high level of expertise required for its interpretation. Nevertheless, it represents a useful new protein identification method which greatly increases the power of peptide mass fingerprinting protein identification.

Cross-species protein identification

Protein sequence databases continue to grow at a rapid rate, yet it is not widely appreciated that close to 90% of all information contained in current protein databases comes from only 10 species (A. Bairoch, Pers. Comm.). Fortunately, this information can be used to study proteomes of organisms that are poorly defined at the molecular level, via 2-D electrophoresis and 'cross-species' protein identification (Cordwell *et al.*, 1995; Wasinger *et al.*, 1995). This approach allows proteins from reference maps of many different species to be identified without the need for the corresponding genes to be cloned and sequenced. This is particularly true for 'housekeeping' proteins, such as enzymes involved in glycolysis, DNA manipulation and protein manufacture, which are highly conserved across species boundaries. Proteins that cannot be identified across species boundaries can then become the focus of further protein characterisation and DNA sequencing efforts.

A)

```
Protein APA1_HUMAN
==================

Asx:  8.4   Glx: 19.3   Ser:  6.3   His:  1.3
Gly:  4.2   Thr:  4.3   Ala:  8.0   Pro:  4.2
Tyr:  2.9   Arg:  6.7   Val:  5.5   Met:  1.3
Ile:  0.0   Leu: 15.5   Phe: 2.5   Lys:  8.8

pI Range: no range specified
Mw Range: no range specified

The closest SWISS-PROT entries are:

Rank Score   Protein    (pI       Mw)  Description
=================================================
   1     0 APA1_HUMAN   5.27    28078  APOLIPOPROTEIN A-I.
   2     4 APA1_MACFA   5.43    28005  APOLIPOPROTEIN A-I.
   3    12 APA1_RABIT   5.15    27836  APOLIPOPROTEIN A-I.
   4    14 APA1_BOVIN   5.36    27549  APOLIPOPROTEIN A-I.
   5    14 APA1_CANFA   5.10    27467  APOLIPOPROTEIN A-I.
   6    18 APA1_MOUSE   5.42    27922  APOLIPOPROTEIN A-I.
   7    26 APA1_PIG     5.19    27598  APOLIPOPROTEIN A-I.
   8    27 APA1_CHICK   5.26    27966  APOLIPOPROTEIN A-I.
   9    37 DYNA_CHICK   5.44   117742  DYNACTIN, 117 KD ISOFORM.
  10    39 APA4_HUMAN   5.18    43374  APOLIPOPROTEIN A-IV.
```

B)

```
Reagent: Trypsin  MW filter: 10%

Scan using fragment mws of:

1953  1933  1731  1613  1401  1387
1301  1283  1252  1235  1231  1215
1031   896   873   831   813   781
 732   704

No. of database entries scanned = 72018

1  . APA1_HUMAN   APOLIPOPROTEIN A-I (APO-AI). - HOMO SAPIENS
2  . APA1_MACFA   APOLIPOPROTEIN A-I (APO-AI). - MACACA FASCICULARIS
3  . APA1_PAPHA   APOLIPOPROTEIN A-I (APO-AI). - PAPIO HAMADRYAS
4  . B41845       orf B - Treponema denticola
5  . APA1_CANFA   APOLIPOPROTEIN A-I (APO-AI). - CANIS FAMILIARIS (DOG).
6  . S30947       hypothetical protein 1 - Azotobacter vinelandii
7  . HS2C_PEA     CHLOROPLAST HEAT SHOCK PROTEIN PRECURSOR. - PISUM SATIVU
8  . S20724       Tropomyosin - African clawed frog
9  . HIVVI354     HIVVI354 premature term. at 793 - Human immunodeficiency
10 . TRJ2_ECOLI   TRAJ PROTEIN. - ESCHERICHIA COLI.
```

Figure 6. Theoretical cross-species matching of human apolipoprotein A-I by amino acid composition and tryptic peptides. When an unknown protein is analysed, best ranking proteins from both techniques can be compared. If the same protein type is observed in both lists, there is high confidence in this being the identity of the unknown molecule (Cordwell *et al.*, 1995). (A) Output of ExPASy server (Appel, Bairoch and Hochstrasser, 1994) where the true amino acid composition of apolipoprotein A-I was matched against all entries in the SWISS-PROT database, without pI or MW windows. Seven of the top 10 matching proteins were apolipoprotein A-I of different species. (B) Output of MOWSE peptide mass fingerprinting program (Pappin, Hojrup and Bleasby, 1993) where true tryptic peptides of human apolipoprotein A-I were matched against the OWL database, using MW window of 10%. Four of the top ten matching proteins were apolipoprotein A-I from different species.

Rapid cross-species identification of proteins from 2-D reference maps can be undertaken with amino acid composition or peptide mass fingerprinting methods (*Figure 6*), but these techniques alone may not identify proteins unambiguously when phylogenetic cross-species distances are great or analysis data is of poor quality (Yates *et al.*, 1993; Shaw, 1993; Cordwell *et al.*, 1995). However, very high confidence in protein identities can be achieved when lists of best-matching proteins generated by both techniques are compared (Cordwell *et al.*, 1995; Wasinger *et al.*, 1995). The correct identification is found when the same protein is ranked highly in lists of best matches generated by both techniques. This method has allowed approximately 120 proteins from the reference map of the mollicute *Spiroplasma melliferum*, representing approximately one quarter of the proteome, to be confidently identified by reference to protein information from other species (S. Cordwell, Personal Communication). When cross-species protein identification is to be undertaken, it should be noted that the molecular weight of a protein type across species is usually highly conserved, but that protein pI can vary by more than 2 units (Cordwell *et al.*, 1995). Accurate molecular weight determination by direct mass spectrometry of proteins blotted to PVDF (Eckerskorn *et al.*, 1992) should therefore be a useful additional parameter for cross-species protein identification.

CHARACTERISATION OF POST-TRANSLATIONAL MODIFICATIONS

Many proteins are modified after translation. Such post-translational modifications, including glycosylation, phosphorylation, and sulfation (see *Table 6*), are usually necessary for protein function or stability. Some abnormal modifications are associated with disease (Duthel and Revol, 1993; Ghosh *et al.*, 1993; Yamashita *et al.*, 1993). In proteome studies, post-translational modifications can be examined on all proteins present, or on individual spots. Studies on all proteins provide an indication of which proteins may carry a certain type of modification. For example, 2-D gel analysis of cell cultures grown in the presence of [^{3}H] mannose or [^{32}P] phosphate gives an indication of which proteins carry glycans containing mannose, and which proteins are phosphorylated (Garrels and Franza, 1989). Lectin binding studies of 2-D gels blotted to PVDF or nitrocellulose provide information on the saccharides, if any, that are carried by proteins present (Gravel *et al.*, 1994).

When individual proteins of interest carrying post-translational modifications have been found, micropreparative 2-D electrophoresis can be used to purify them in microgram quantities (Hanash *et al.*, 1991; Bjellqvist *et al.*, 1993b). If protein isoforms of similar MW and pI are to be studied, focusing with narrow range pI gradients (1 pH unit) can provide greater separation and resolution. After electrophoresis, the type and degree of protein phosphorylation can be investigated (Murthy and Iqbal, 1991; Gold *et al.*, 1994), monosaccharide composition can be determined (Weitzhandler *et al.*, 1993; Packer *et al.*, 1995), and the structure and exact site of glycoamino acids can be investigated by either Edman degradation based techniques or by mass spectrometry (Pisano *et al.*, 1993; Huberty *et al.*, 1993; Carr, Huddleston and Bean, 1993). With further development of rapid techniques, investigation of phosphorylation and monosaccharides by chromatographic or mass spectrometric means is likely to become a routine step in the characterisation of post-translational modifications of proteins from reference maps.

The status of proteome projects

Many technical aspects of proteome research have already been discussed in this review, but an overview of the status of proteome projects has not yet been presented. Advances in proteome projects will initially rely on progress in genome sequencing initiatives, to enable an identity, amino acid sequence, or function to be assigned to each protein spot. *Table 7* shows genome size, proteome size, and the number of proteins already defined for a number of model organisms. This indicates that whilst genome sequencing programs for *E. coli* and *S. cerevisiae* are advanced, the massive size of some other genomes (and especially the human genome) means that their complete nucleotide sequences are unlikely to be available for many years. Because of this, 2-D reference maps and proteome projects of single cell organisms like *Mycoplasma sp.*, *E. coli* and *S. cerevisiae* will be the most detailed (Cordwell *et al.*, 1995; Wasinger *et al.*, 1995; Vanbogelen *et al.*, 1992; Garrels *et al.*, 1994), and complete maps of other organisms will take longer to construct. However, the use of cross-species protein identification techniques will allow proteomes of many prokaryotes and simple eukaryotes to be partially defined in reference to *E. coli* and *S. cerevisiae*.

Table 7: Estimated genome size, estimated proteome size, number of protein sequences in SWISS-PROT Release 31 (March, 1995), and approximate number of proteins of known identity on 2-D reference maps for some model organisms. Genome size data from Smith (1994), and total protein data from Bird (1995). Genome sequencing projects of *E. coli* and *S. cerevisiae* will probably be complete in 1996.

Species Name	Haploid genomesSize (million bp)	Estimated proteome size (total proteins)	Protein entries in SWISS PROT	Proteins annotated on 2-D Maps
Mycoplasma species	0.6–0.8	400–600	100	> 100
Escherichia coli	4.8	4000	3170	> 300
Saccharomyces cerevisiae	13.5	6000	3160	> 100
Dictyostelium discoideum	70	12500	204	–
Arabidopsis thaliana	70	14000	270	–
Caenorhabditis elegans	80	17800	703	–
Homo sapiens	2900	60000–80000	3326	> 1000

The study of vertebrate proteomes and vertebrate development is a phenomenal undertaking in comparison to the investigation of single cell organisms. This is because vast numbers of proteins are developmentally expressed, each body tissue has hundreds of unique proteins, and there are numerous tissue types. However, it is estimated that at least 35% of proteins in vertebrate cells will be conserved from tissue to tissue, constituting the 'housekeeping' proteins (Bird, 1995), with the remainder of proteins constituting a set that are specific to a cell type. Providing that standardised electrophoretic conditions are used, reference maps from many tissues of one organism can be superimposed in gel databases (e.g. Hochstrasser *et al.*, 1992). This accelerates the definition of the 'housekeeping' proteins, as well as sets of proteins that are unique to different tissue types. Such studies may, however, be complicated by post-translational modifications, which can differ on the same gene product in different tissues. Proteins that remain unknown after identification procedures will be useful in providing focus for nucleic acid sequencing initiatives.

FUTURE DIRECTIONS OF PROTEOME PROJECTS

This review has described recent advances in the area of proteome research. It has illustrated how new developments of older techniques (2-D electrophoresis and amino acid analysis) as well as the applications of new technology (mass spectrometry) have greatly widened the choice of tools the biologist and protein chemist has for the separation, identification and analysis of complex mixtures of proteins. This has made possible the establishment of detailed reference maps for organisms, which are becoming the method of choice for the definition of tissues or whole cells, and the investigation of gene expression therein.

Proteome projects are already impacting on the dogma of molecular biology that DNA sequence constitutes the definition of an organism. For example, the proteomes of different tissues of a single organism are often significantly different. Similarly, cross-species identification of proteins (for example the identification of proteins from *Candida albicans* by comparison with *S. cerevisiae*) can open up studies on organisms that are poorly molecularly defined. As cross-species identification can proceed at a pace orders of magnitude faster than a genome project in terms of defining the gene and protein complement of organims, the need for the DNA sequencing of genomes will be avoided, and emphasis placed on those found to be novel.

Just as genome sequencing is not an end in itself, neither is an annotated 2-D protein reference map of an organism, nor indeed the identification of proteins in a proteome. So whilst an immediate aim of proteome projects is to screen proteins in reference maps, this will lead to expression studies and characterisation of post-translational modifications. The challenge that then needs to be addressed is the investigation of structure and function of proteins in a proteome. The magnitude of this is illustrated by the fact that over half the open reading frames identified in *S. cerevisiae* chromosome III were initially of no known function (Oliver *et al.*, 1992). Structural and functional studies will be an undertaking just as formidable as genome studies are now and proteome projects are becoming, but will lead to an unimaginably detailed understanding of how living organisms are constructed and how they operate.

Acknowledgements

MRW is the recipient of an Australian Postgraduate Research Award. AAG, MRW, IHS and KLW acknowledge assistance for proteome projects through Macquarie University Research Grants, the Australian Research Council, the Australian National Health and Medical Research Council, Beckman Instruments and GBC Scientific Equipment. DH acknowledges the financial support of a Montus Foundation Grant and the Swiss National Fund for Scientific Research (Grant # 31-33658.92). We thank colleagues who supplied work that was 'In Press' during the writing of this review.

References

ANDERSON, N.L., HOFMANN, J.P., GEMMELL, A. AND TAYLOR, J. (1984). Global approaches to quantitative analysis of gene-expression patterns observed by use of two-dimensional gel electrophoresis. *Clinical Chemistry*, **30**, 2031–2036.